Assignment 03
(due on 11/05 19:00)

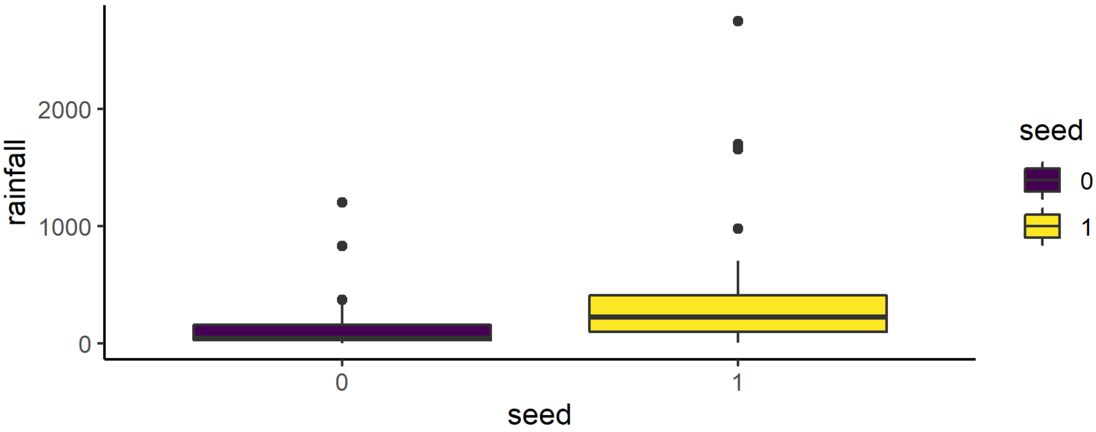
PS3_1.R

1. Cloud Seeding

1.1 [5 points] Plot two box plots side-by-side of the data. Discribe the distributions.

Answer:

```
Rainfall_Plot <- ggplot(Rainfall_data_tbl, aes(x = seed, y = rainfall, fill = seed)) +
  geom_boxplot() +
  theme_classic()
ggsave("Rainfall_plot.png",Rainfall_Plot,width = 15,height = 6,units = "cm")
```



1.2 [5 points] Did cloud seeding have an effect on rainfall in this experiment? If so, how much?

Answer:

```
  seed  count mean_rainfall sd_rainfall
  <ord> <int>         <dbl>       <dbl>
1 0        26          165.        278.
2 1        26          442.        651.
```

```
> summary(anova_one_way)
            Df   Sum Sq Mean Sq F value Pr(>F)
seed         1  1000360 1000360   3.993 0.0511 .
Residuals   50 12525457  250509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

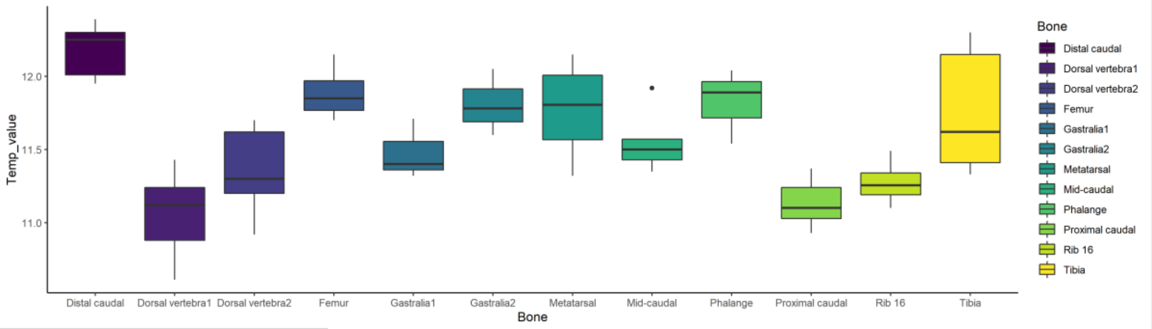
As the result from the anova summary, the P-value is 0.0511, the cloud seeding have little effect on rainfall.

PS3_2.R

2. Was Tyrannosaurus Rex Warm-Blooded?

[10 points] Is there evidence that the means are different for the different bones? Does the dataset support Tyrannosaurus Rex is warm-blooded or not?

Answer:



```
> summary(anova_one_way)
            Df Sum Sq Mean Sq F value   Pr(>F)    
Bone        11  6.067  0.5516   7.427 9.73e-07 ***
Residuals   40  2.971  0.0743                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the result from the anova summary, the P-value is 9.73e-07, the h0 hypothesis should be refused, so the means are different for the different bones. Then the dataset support Tyrannosaurus Rex is warm-blooded.

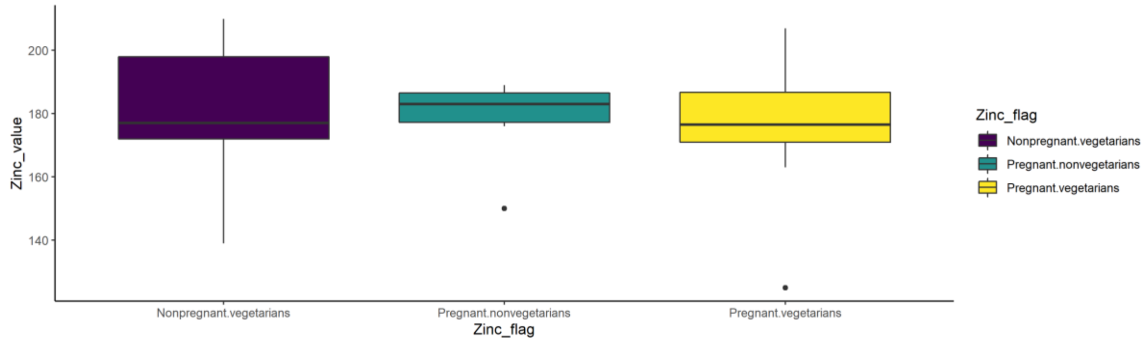
PS3_3.R

3. Vegetarians and Zinc

[10 points] What evidence is there that pregnant vegetarians tend to have lower zinc levels than pregnant nonvegetarians?

Answer:

```
t.test(Zinc_pregnant_veg,na.omit(Zinc_pregnant_nonveg))
	Welch Two Sample t-test
data:  Zinc_pregnant_veg and na.omit(Zinc_pregnant_nonveg)
t = -0.1086, df = 13.947, p-value = 0.9151
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -19.02637  17.19303
sample estimates:
mean of x mean of y 
 177.0833  178.0000 
```



```
summary(anova_one_way)
        Df Sum Sq Mean Sq F value Pr(>F)
Zinc_flag   2     16     8.1   0.018  0.982
Residuals   20   8816   440.8
```

(1)As the result from the t-test, the the P-value is 0.9151,
(2)and as from the anova summary, the P-value is 0.982, the h0 hypothesis should be accepted, so there isn’t any evidence that pregnant vegetarians tend to have lower zinc levels than pregnant nonvegetarians.

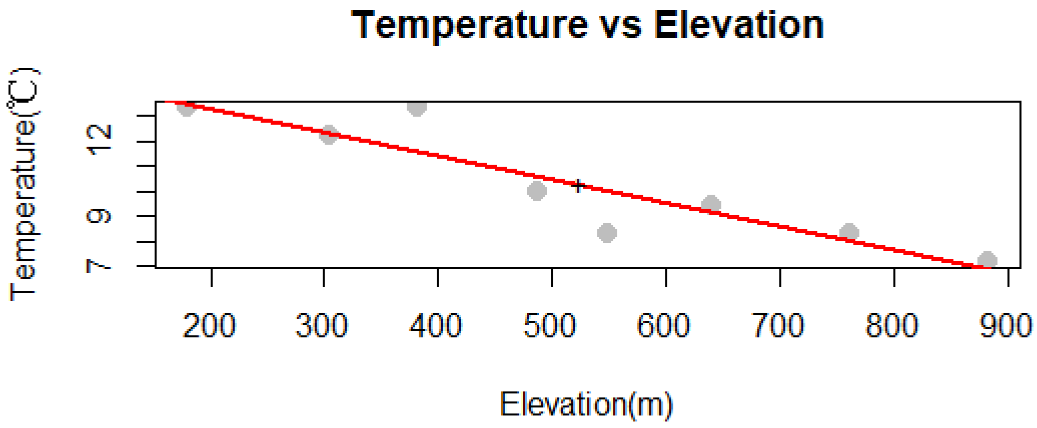
PS3_4.R

4. Atmospheric lapse rate

[10 points] Draw a scatter plot with regression line, and investigate if the lapse rate is 9.8 degrees C km^{-1}.

Answer:

```
             Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 15.124886623 0.948282001 15.949777 3.856494e-06
x          -0.009312104 0.001669811 -5.576742 1.410783e-03
```



The result shows that the lapse rate is 9.3 degrees C km^{-1} with the surveys. It’s a bit different from the given data 9.8.

PS3_5.R

5. The Big Bang Theory

5.1 [5 points] Make a scatter plot with distance as the Y-axis and recession velocity as the x-axis. Describe what you see.

Answer:

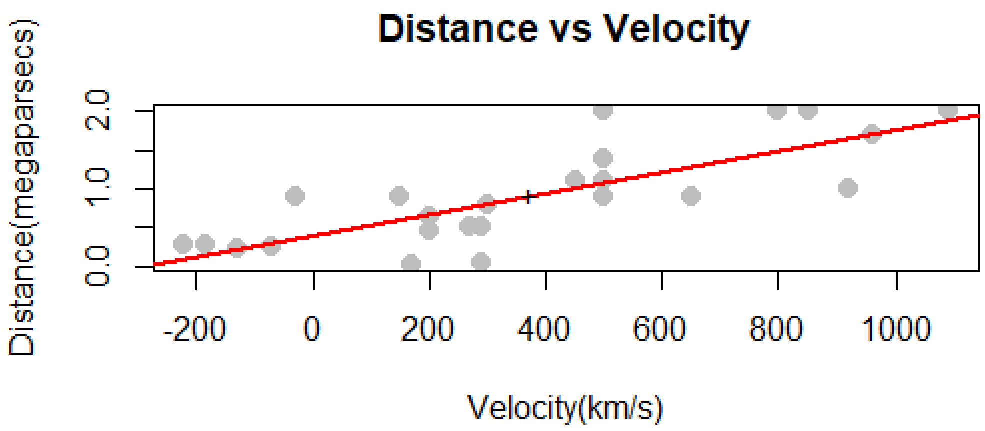
```
plot(y ~ x,
    xlab = "Velocity(km/s)",
    ylab = "Distance(megaparsecs)",
    main = "Distance vs Velocity",
    pch = 20,
    cex = 2,
    col = "grey")
```

I see that the larger the velocity is, the further the distance it is.

5.2 [5 points] Add a simple linear regression to the above scatter plot.

Answer:

```
fit <- lm(y ~ x)
abline(fit, lwd = 2, col = "red")
```



5.3 [10 points] If Hubble's Big Bang theory is correct, then for the regression line:

- The intercept is zero - as the universe starts with one single point!
- And the slope is the age of the universe.

Address the above two assumptions with the dataset and the regression results; and estimate the age of the universe.

Answer:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3990982  0.1184697   3.369  0.00277 **
x           0.0013729  0.0002274   6.036 4.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.405 on 22 degrees of freedom
Multiple R-squared:  0.6235,        Adjusted R-squared:  0.6064
F-statistic: 36.44 on 1 and 22 DF,  p-value: 4.477e-06
```

(1)As the p-value of the intercept is 0.00277 **, and the p-value of the slope is 4.48e-06 *, the Big Bang theory could be right.**
(2) According to the Big Bang theory , the slope is the age of the universe, then the age of the universe could be estimated by unit conversion.
For 1mpc=$1*10^{6}*30.9*10^{12}$km, then
Y= 0.0013729 *[($1*10^{6}*30.9*10^{12}$km) / 1(km/s)] /(365.25*24*60*60)
= 0.0013729 *($3.09*10^{19}$s)/ (365.25*24*60*60)
= 1.344 *10^{9}a
=1.344billion year.

5.4 [5 points] Explain why improved measurement of distance would lead to more precise estimates of the regression coefficients.

Answer:

There is a linear relationship between the redshift and the distance of the galaxy spectrum. The farther away the galaxy is, the higher the redshift is. According to the Doppler effect, the redshift of the spectrum of galaxies is due to the fact that the galaxies are moving away from the Milky way. The higher the redshift value, the faster the star system will regress.

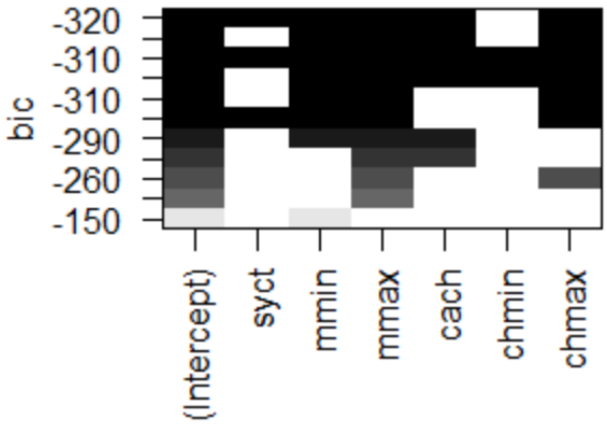
PS3_6.R

6. CPU performance

6.1 [5 points] For your train set, fit a best subset regression between predictor variable perf and response variables syct, mmin, mmax, cach, chmin, chmax.

Answer:

```
cpus_train_result <- regsubsets(perf ~ syct+ mmin + mmax + cach +
                             chmin + chmax, data=cpus_train, nbest=2, nvmax = 6)
plot(cpus_train_result, scale="bic")
```



```
# Build a linear model
fullmodel=lm(perf ~ syct+ mmin + mmax + cach + chmin + chmax, data=cpus_train)
model_step_b <- step(fullmodel,direction='backward')

# Get estimates
summary(fullmodel)
```

```
> summary(fullmodel)
Call:
lm(formula = perf ~ syct + mmin + mmax + cach + chmin + chmax,
    data = cpus_train)

Residuals:
    Min      1Q  Median      3Q     Max
-194.21  -29.22    2.87   28.29  341.77

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.933e+01  8.942e+00  -6.635 4.76e-10 ***
syct         4.983e-02  1.949e-02   2.556   0.0115 *
mmin         1.417e-02  2.069e-03   6.850 1.51e-10 ***
mmax         6.574e-03  7.478e-04   8.792 2.21e-15 ***
cach         8.003e-01  1.742e-01   4.595 8.74e-06 ***
chmin       -8.474e-01  9.538e-01  -0.888   0.3756
chmax        1.362e+00  2.427e-01   5.612 8.61e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.2 on 160 degrees of freedom
Multiple R-squared:  0.8775,	Adjusted R-squared:  0.873
F-statistic: 191.1 on 6 and 160 DF,  p-value: < 2.2e-16
```

6.2 [10 points] Apply the best regression model to the test set, and compare your predicted perf with the actual values provided. Quantify the bias between predicted perf values and provided perf values.

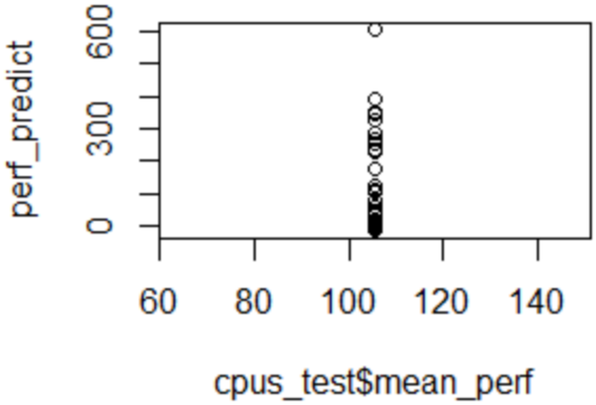
Answer:

```
# Apply the model model_log to the test subset
```

```
perf_predict <- predict(fullmodel,cpus_test)

# Compare predicted values with actual values
plot(cpus_test$mean_perf, perf_predict)

# Relative mean bias
(mean(perf_predict) - mean(cpus_test$mean_perf))/mean(cpus_test$mean_perf)
```



The bias between predicted perf values and provided perf values is 0.1668544.

7. Analysis of a data set

Find some data sets from your research group, conduct the following statistical tests:

7.1 [5 points] Define a simple research question that can be tested with the t-test. Test your question with R, and describe your findings.

Answer:

There is a water monitoring station at the A-river. It puts out daily water temperature and PH value everyday. We have some data from August to October, and we want to know is there evidence that the PH value means are different between August and September.

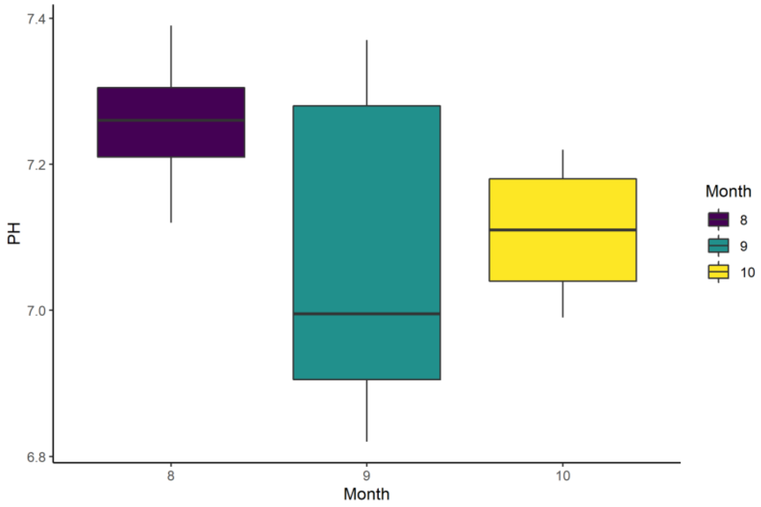
```
t.test(Monitoring_Aug$PH, Monitoring_Sep$PH)
            Welch Two Sample t-test
data:  Monitoring_Aug$PH and Monitoring_Sep$PH
t = 5.2946, df = 35.921, p-value = 6.128e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1219975 0.2735079
sample estimates:
mean of x mean of y
 7.257419  7.059667
```

As the result from t-test, the P-value is 6.128e-06, so the answer is that the PH value means are different between August and September.

7.2 [5 points] Define a simple research question that can be tested with the ANOVA. Test your question with R, and describe your findings.

Answer:

The data is the same as 7.1, and we want to know is there evidence that the means are different for the different months.



```
summary(anova_one_way)
            Df Sum Sq Mean Sq F value   Pr(>F)    
Month        2 0.6366  0.3183   18.29 3.17e-07 ***
Residuals   77 1.3401  0.0174                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the result from one way anova test, the P-value is 3.17e-07, so the answer is that the PH value means are different for the different months.

7.3 [5 points] Define a simple research question that can be tested with a simple linear regression model. Test your question with R, and describe your findings.

Answer:

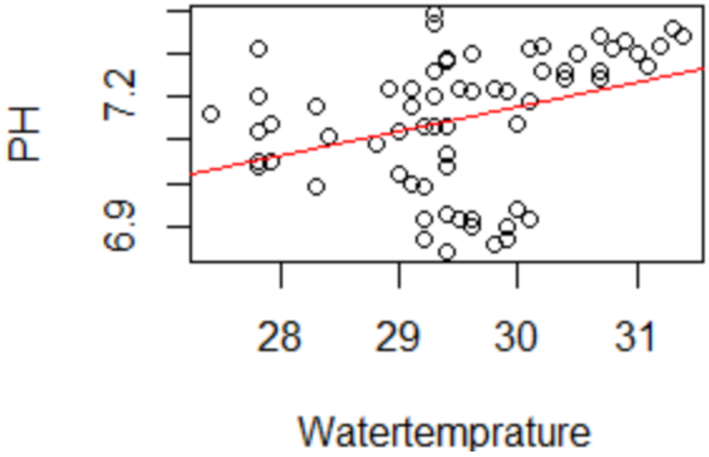
The data is the same as 7.1, and we want to fit the best subset regression between predictor variable **PH** and response variable **Watertemprature**.

```
Call:
lm(formula = PH ~ Watertemprature, data = Monitoring_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.33904 -0.07588  0.05248  0.10100  0.25662 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.63006    0.55281  10.184 3.61e-15 ***
Watertemprature  0.05131    0.01875   2.737  0.00797 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.1494 on 66 degrees of freedom
Multiple R-squared:  0.1019,	Adjusted R-squared:  0.0883 
F-statistic: 7.489 on 1 and 66 DF,  p-value: 0.007968
```



As the result from linear regression model, the Adjusted R-squared: is 0.0883, the result is not good, so perhaps the PH value has not much relationship with Watertemprature.